Sampling strategy, characteristics and representativeness of the InGef research database

Marion Ludwig
InGef - Institute for Applied Health Research Berlin GmbH https://orcid.org/0000-0002-2446-0090
Dirk Enders
InGef - Institute for Applied Health Research Berlin GmbH
Frederike Basedow
InGef - Institute for Applied Health Research Berlin GmbH
Jochen Walker
InGef - Institute for Applied Health Research Berlin GmbH https://orcid.org/0000-0002-7838-7829
Josephine Jacob (josephine.jacob@ingef.de)
InGef - Institute for Applied Health Research Berlin GmbH https://orcid.org/0000-0001-8769-9747

Research Article

Keywords: data sources, healthcare databases, claims data, external validity, pharmacoepidemiology

Posted Date: February 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1046019/v2

Version of Record: A version of this preprint was published at Public Health on May 1st, 2022. See the published version at https://doi.org/10.1016/j.puhe.2022.02.013.

Abstract

Objectives

The aim of this study was to describe the sampling strategy as well as characteristics and the external validity of a representative sample database drawn from the German InGef research database.

Study Design

This is a retrospective cohort study using anonymized claims data for the year 2019.

Methods

The InGef research database is an anonymized healthcare database with longitudinal claims data from approximately 8.8 Mio insurees. A sample of four million insurees was drawn intended to be representative for the German population with respect to age, sex and region. In addition to demographic information, data on hospitalization rates, mortality rates and drug prescription rates was analysed from the InGef sample database for the year 2019 to demonstrate validity and representativeness. Corresponding national reference data were obtained from official sources.

Results

The distributions of sex and age were similar in the InGef sample database and Germany (proportion of women: 50.8% vs. 50.7%; mean age: 44.1 vs 43.9 years). The proportion of insurees living in the eastern part of Germany was lower in the InGef sample database (16.5% vs. 19.5 %). There was good accordance with German reference data with respect to hospitalization rates and overall mortality rates. Prescription rates for the 20 most often reimbursed drug classes were slightly higher in the InGef sample database.

Conclusions

The InGef sample database shows good overall agreement with the German population on measures of morbidity, mortality, and drug usage.

Background

Using claims data for health service research purposes has steadily increased in Germany over the last years, showing that routine data are becoming a more common and important source for health services research (1–4). Claims data (insurance data) is routinely collected for administration and reimbursement purposes. In Germany, about 85% of German inhabitants are covered by statutory health insurances (SHI). These data offer enormous potential for health services research, including health economic or pharmacoepidemiological studies. As such, they provide cross-sector data content that is, unlike primary data or survey data, free of selection or recall bias. Further, claims data are available in a large sample size and offer the possibility of longitudinal analyses. In contrast to comprehensive all payer claims databases (APCD) found in the US or Canada (1,5,6), various databases exist in Germany, which allow only very limited access and are still underreported regarding their content and validity. The InGef research database consists of anonymized data from approximately 8.8 Mio individuals, who are insured with one of the 58 German SHIs currently contributing data to the database. A sample database (InGef sample database) consisting of 5% of the German population (~four million insurees) is drawn to ensure representativeness and validity of the database used for health services research. The aim of this study is to describe the sampling method of the representative InGef sample database and to demonstrate its representativeness for the German population based on relevant demographic and clinical measures.

Methods

Data source

The InGef research database currently (December 2020) includes longitudinal data of approximately 8.8 Mio SHI members, insured in one of the contributing SHIs (mainly company or guild health insurances), and covers insurees from all federal states of Germany. The claims data are collected in a specialized data centre owned by SHIs, providing data warehouse and IT services. Data are anonymized before entering the InGef research database by the data centre, acting as a trust centre for this anonymization process. The anonymization process ensures that an identification of insured individuals, health care providers, and the respective SHI is not possible. Moreover, access to the InGef research database data is strictly controlled as well as project bound, and analyses are performed exclusively by InGef employees. The lag time of data availability is about nine months. Sampling of the InGef sample database starts with the year 2011. However, due to data privacy regulations for research projects, the time period covered by the InGef sample database is limited to a 6-year look-back (starting with the most recent complete data year).

German SHI claims data available in the InGef research database includes partly coarsened information on demographics (quarter of birth, sex, quarter of death if applicable, region of residence on administrative district level); inpatient care (diagnoses, diagnosis related groups (DRG), operation and procedures (OPS) (7)); outpatient services (diagnoses, treatments, specialities of physicians); dispensing of reimbursed drugs; dispensing of reimbursed remedies,devices and aids; and sick leave and sickness allowance times. In addition, costs from the SHI perspective are available of all healthcare sectors. Diagnoses in the InGef database are coded using the International Classification of Diseases, version 10 in the German Modification (ICD-10-GM) (8). Prescriptions of medication are identified based on Anatomical Therapeutic Chemical (ATC)-codes as classified by the AOK Research Institute (WIdO) (9).

Sampling Strategy

The aim of the sampling strategy for the InGef research database was to select a representative sample (InGef sample database) of the German population with respect to age, sex and, of minor priority, region of residence, which allows studies on various research questions in health services research including rare diseases or complex treatment patterns. Therefore, a sample of 5% sample of the German population was drawn annually (~ four million insurees) to continuously ensure representativeness and validity (Figure 1). The sampling strategy favoured persons with complete data, i.e. valid data from all data strands (including insurance time, incapacity to work and remedies and aids).

The desired properties of the InGef sample database had the following general consequences on the sampling strategy. First, the sample size required to draw a given number of insurees in age and sex strata of the InGef research database. If there were not enough insurees in the InGef research database for a stratum, insurees of a close stratum had to be drawn instead. Second, insurees leave the InGef research database if they die or change the health insurance. Due to the annual sampling procedure, newborns for each sampling year can initially not be included in the InGef research database, as they were not available in the year before. Accordingly, all newborns must be drawn in the annual sampling from the pool of insured persons. The German population also changes because of migration, death, or birth. Therefore, the InGef sample database had to be adapted each year to preserve representativity over the years (Figure 1).

The reference population for the InGef sample database was 5% of the German population in the categories age (0-89 in yearly steps, 90+ years), sex and federal state, based on the statistics of the Federal Statistical Office (10) (further tables were provided upon request by the Federal Statistical Office). Federal state is further coarsened into North (Hamburg, Bremen, Schleswig Holstein, Lower Saxony, Mecklenburg Western Pomerania), South (Baden-Württemberg, Bavaria), West (Northrhine-Westphalia, Hesse, Rhineland Palatinate, Saarland) and East (Berlin, Brandenburg, Saxony, Saxony-Anhalt, Thuringia).

Sampling started with the year 2011 and was performed as follows:

1. The reference population at 31st Dec 2011 was extracted, cross-tabulated by age, sex and region of residence. All persons in the InGef research database at 31st Dec 2011 were eligible for sampling except those with data anomalies (i.e.: more than one birthday, insured after death or before birth). Sex and region of residence were determined on the last available date for each insuree in the database in the same categories as in the reference population. Age was calculated as of 31st Dec in each year. The normalized insurance time was defined as the fraction of insured time over three consecutive years, i.e.: insured time from 1st Jan 2012 to 31st Dec 2014 divided by the total time from 1st Jan 2011 to 31st Dec 2013. The insurance time is used in the further sampling strategy to prefer insured persons who are available in the database over a long time. This is important since insurance coverage is usually an inclusion criterion for further (longitudinal) analyses. Thus, including insurees with a high normalized insurance time close to one enhances the chance to achieve representativeness even after insurance verification. Normalization was performed for a technical reason to simplify the construction of the sampling algorithm. To avoid underrepresentation of deceased patients, their normalized insurance time was sampled from non-deceased patients.

2. Sampling for the first year (2011) was performed for each age and region of residence stratum in the reference population, separately for males and females. The sampling started with the least represented stratum, i.e.: the stratum with the worst ratio of available insurees in the InGef research database and the required number of persons for that stratum as of the reference population. Insurees were prioritized for sampling, if their age and region of residence agreed with the reference population and if they had complete data, i.e. maximal normalized insurance time and complete information on incapacity of work and remedies and aids. Insurees not fulfilling these criteria were assigned a lower priority.

Insurees were drawn according to their priority until the number of individuals required for the stratum was reached. In the same manner, the procedure iterated through the remaining strata until all strata were sampled. Sampled insurees were removed for further iterations.

3. For the sampling in each of the following years (2012-2019), the differences between the reference population for this year and the sample in all strata were determined. Missing persons in the strata were filled as in step 2. To keep representativity of the preceding years, newly drawn insurees, who have been insured in the database in the preceding years but did not belong to the database sample so far, were included at the beginning of one of the quarters of the respective year. The quarters in which observation time begins for newly drawn insurees in a respective calendar year were randomly assigned based on the distribution of all quarters in which all insurees entered the database in that year. All data of the respective insured person before this sampled quarter were not used for the InGef sample database.

Analyses

The following information was extracted: i) demographic information (sex, age, region of residence) of all insurees alive at 31 December 2018 or born in 2019 who were fully insured until 31st December, 2019 or until their date of death in 2019; ii) hospitalization rates grouped by discharge diagnosis (main ICD-10-GM chapters) in 2019; iii) mortality rates in 2019 and iv) drug prescription rates in 2019 (20 most frequently prescribed ATC-groups (2nd level) as number of prescribed packages). The hospitalization rate per defined ICD-chapter was calculated by dividing the number of hospitalizations (fully inpatient) with a discharge date between 01.01.2019 – 31.12.2019 and a main discharge diagnosis of a respective ICD-chapter by the total number of fully insured persons in the InGef sample database in 2019. The drug prescription rate was calculated as the sum of the quantity of packages prescribed of all insured persons divided by the total number of fully insured persons in the InGef sample database in 2019. The drug prescription rate for the German reference population was calculated accordingly. The mortality rate was calculated by dividing the number of deceased insurees by the number of fully insured persons in the InGef sample database in 2019.

For hospitalization rates, national reference data for 2019 were extracted from the Information System of the Federal Health Monitoring (11) and in alignment with the Federal Statiscial Office (Destatis) data (12) for the total German population. For hospitalization rates diagnosis data based on the place of treatment were used, since the InGef sample database includes person with a residence abroad. Mortality rate of the total German population was extracted from the Federal Health Monitoring (13). National reference data for the distribution of age, sex and region of residence was extracted from the Federal Statistical

Office (14). Drug prescriptions for the German population insured within the SHI system was taken from the German Drug Prescription Report 2020 (15) and the Federal Health Monitoring System (16).

The mean continuous insurance time in the InGef sample database was determined as the time from Jan. 1, 2014 or entry into the database, to the end of insurance, death, or Dec. 31. 2019, whichever came first, in years.

Results

Mean age of insurees in the InGef sample database was in good accordance with the German population (mean age: 44.1 vs 43.9 years). Moreover, the proportion of women in the InGef sample database corresponded well to the proportion in the total German population (50.8% vs. 50.7%, InGef database vs. German population). The percentage of insurees living in the Eastern parts of Germany and the proportion of persons living in rural areas was slightly lower in the InGef sample database compared to the total German population. Table 1 displays the comparison of the main demographic characteristics.The mean continuous insurance time since entry into the InGef sample database was 4.78 ± 2.01 years. The proportion of newly drawn insurees in the year 2019 was 3.33%.Hospitalization rates, mortality rates and drug usage of persons in the InGef database were similar to the German reference data. Hospitalization rates were slightly lower in most main ICD-chapters. Larger deviations were found for Pregnancy, childbirth and the puerperium - ICD-chapter O (InGef vs. Germany: 19.3 vs. 24.5 per 1000 persons) and Certain conditions originating in the perinatal period - ICD-chapter P (InGef vs. Germany: 1.7 vs. 2.4 per 1000 persons) (Figure 2). Out of the 20 most frequently prescribed ATC drug classes, prescription rates for 18 drug classes were slightly higher in the InGef sample database compared to reference data from the German drug prescription report 2020 (15) (Figure 3). The mortality rate of the persons insured in the InGef sample database was slightly lower than in the German population (10.5 vs. 11.3 per 1,000 persons).

Discussion

The InGef sample database demonstrates good overall accordance with the German reference population. Especially, differences in sex and age distribution as well as mortality rates between the InGef sample database and Germany were small. It was previously reported that substantial differences exist in the characteristics and socio-economic standards of the persons insured with the different German SHIs (17–19). Moreover, studies that have examined socio-economic inequalities worldwide, warn that these translate into differences in mortality and morbidity rates (20). Accordingly, lower mortality rates were reported for the year 2006 for the German Pharmacoepidemiological Research Database (GePaRD), a database with a population of presumably higher socio-economic status than the overall German population (21). However, for Germany, van Raalte et al. recently described that regional disparities in mortality based on the large economic inequalities between the German federal states are declining (22), a finding that is supported by our comparative analysis.

Due to the structure of the SHIs that provide data for the InGef research database, the proportion of insurees in the regions of East and West Germany deviates slightly from the German reference population. The differences observed for the hospitalization rates and the prescribed ATC drug classes might likewise be explained by the described regional and socio-economic variations between insurees of the InGef research database and the German population (17,18,23). Especially, the lower hospitalization rates found for ICD-chapters O (Pregnancy, childbirth and the puerperium) and P (Certain conditions originating in the perinatal period) might be linked to a higher socioeconomic status, which has been reported to result in reduced fertility (24,25). Therefore, in observational studies which aim at examining regional differences for specific outcomes, additional standardization with respect to region of residence, should be considered.

External validity is a feature of the InGef sample database and of upmost importance for epidemiological studies comparing the effect of treatments or health interventions. Thus, unless a very high external validity is explicitly required, the observed marginal differences between the InGef sample database and the reference data are neglectable.

Some of the fundamental advantages of claims data characterized here should be mentioned. These secondary data are free of selection and recall bias and contain complete data offering an intersectoral perspective. Some further strengths of the data are the possibility to precisely determine the base population, the large sample sizes and the continuous data collection allowing the monitoring of the state of health over a long time period (26,27). In addition to the known strengths of claims data, the InGef sample database provides a readily available, reliable, and representative data source for healthcare research.

Limitations

The comparison of the InGef sample database with the external reference data showed good accordance. However, although the sampling strategy was designed to draw a sample which is representative for the German population in each year, the representativity might be lower after applying study specific selection criteria. Especially, studies on incident drug use in a given year would not include newly drawn insurees in that year as these persons did not contribute to the database with insurance data in the previous year. However, for the analyzed year 2019 for example, the impact of excluding the newly drawn individuals on the representativity is rather low (3.33% newly drawn insurees).

Further, there are a few limitations that are inherent to claims data or come with the use of the database. First, due to the anonymized nature of the data it is not possible to validate the data using medical charts. Second, data availability for health services research purposes is limited to six years, which is critical for studies that require a longer observation period.

Conclusions

The InGef sample database can be considered representative for the German population and is thus a valuable data source for health services research.

Abbreviations

ATC - Anatomic Therapeutic Chemical Classification Systems

ICD - International Classification of Disease and Health Related Problems, German Modification

OPS - German procedure classification (“Operationen- und Prozedurenschlüssel”)

SHI - Statutory Health Insurance

Declarations

Conflict of Interest

Authors declare no conflict of interest.

Ethic statement

Anonymized SHI data were used for this study. The use of such data is not subject to ethics committee approval in Germany.

Data availability

The data used in this study cannot be made available in the manuscript, the supplemental files, or in a public repository due to German data protection laws (Bundesdatenschutzgesetz). To facilitate the replication of results, anonymized data used for this study are stored on a secure drive at the InGef - Institute for Applied Health Research Berlin GmbH. Access to the raw data used in this study can only be provided to external parties under the conditions of a cooperation contract and can be accessed upon request, after written approval (info@ingef.de), if required.

Funding

The project did not utilize any funding.

References

1. Kreis K, Neubauer S, Klora M, Lange A, Zeidler J: Status and perspectives of claims data analyses in Germany-A systematic review. Health Policy Amst Neth 2016; 120: 213–26.

2. Neubauer S, Zeidler J, Schilling T, et al.: Suitability and Usability of Claims Data for Review of Guidelines for the Treatment of Chronic Heart Failure. Gesundheitswesen 2016; 78: e135–44.

3. Gansen FM: Health economic evaluations based on routine data in Germany: a systematic review. BMC Health Serv Res 2018; 18: 268.

4. Neubauer S, Kreis K, Klora M, Zeidler J: Access, use, and challenges of claims data analyses in Germany. Eur J Health Econ 2017; 18: 533–6.

5. Love D, Custer W, Miller P: All-payer claims databases: state initiatives to improve health care transparency. Issue Brief Commonw Fund 2010; 99: 1–14.

6. APCD Council. APCD Council. https://www.apcdcouncil.org/ (last accessed on 2021 Aug 25)

7. Deutsches Institut für Medizinische Dokumentation und Informatik: Operationen- und Prozedurenschlüssel Version 2020. 2020. https://www.dimdi.de/static/de/klassifikationen/ops/kode-suche/opshtml2020/ (last accessed on 2020 Jul 6)

8. Deutsches Institut für Medizinische Dokumentation und Informatik: Internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme 10. Revision German Modification Version 2020. 2020. https://www.dimdi.de/static/de/klassifikationen/icd/icd-10-gm/kode-suche/htmlgm2020/ (last accessed on 2020 Jul 6)

9. ATC-Klassifikation. https://www.dimdi.de/dynamic/de/arzneimittel/atc-klassifikation (last accessed on 2021 Aug 19)

10. Statistisches Bundesamt Deutschland - 12411-0013: Bevölkerung: Bundesländer, Stichtag, Geschlecht, Altersjahre. GENESIS-ONLINE. 2021. https://www-genesis.destatis.de/genesis//online?operation=table&code=12411-0013&bypass=true&levelindex=0&levelid=1617953251928#abreadcrumb (last accessed on 2021 Apr 9)

11. Diagnosedaten der Krankenhäuser 2019. https://www.gbe-bund.de/gbe/!pkg_olap_tables.prc_set_page?p_uid=gast&p_aid=41335358&p_sprache=D&p_help=2&p_indnr=550&p_ansnr=99818202&p_version=10&D.001=1000001&D.946=1000468&D.011=44302 (last accessed on 2022 Jan 12)

12. Destatis Vollstationäre Patientinnen und Patienten der Krankenhäuser. Statistisches Bundesamt. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Gesundheit/Krankenhaeuser/Tabellen/diagnose-kapitel-geschlecht.html (last accessed on 2022 Jan 12)

13. Mortalität - Sterbefälle Sterbeziffern ab 1998. https://www.gbe-bund.de/gbe/!pkg_olap_tables.prc_set_page?p_uid=gast&p_aid=41335358&p_sprache=D&p_help=2&p_indnr=6&p_ansnr=85247890&p_version=2&D.000=3741&D.001=1000001&D.002=1000002&D.003=1

(last accessed on 2022 Jan 12)

14. Statistisches Bundesamt Deutschland - GENESIS-Online Alter Geschlecht. 2022. https://www-genesis.destatis.de/genesis/online?operation=abruftabelleBearbeiten&levelindex=0&levelid=1641982617227&auswahloperation=abruftabelleAuspraegungAuswaehlen&auswahlverzeichnis=ordn0013&auswahltext=&werteabruf=Werteabruf#abreadcrumb (last accessed on 2022 Jan 12)

15. Schwabe U, Ludwig W-D (eds.): Arzneiverordnungs-Report 2020. Berlin Heidelberg: Springer-Verlag 2020.

16. Therapeutische Arzneimittel zu Lasten der GKV 2019. https://www.gbe-bund.de/gbe/pkg_olap_tables.prc_sort_ind?p_uid=gast&p_aid=84612583&p_sprache=D&p_help=2&p_indnr=613&p_ansnr=32092685&p_version=4&p_sortorder=d&p_dim_1=D.100&p_dw_1=10101 (last accessed on 2022 Jan 6)

17. Hoffmann F, Icks A: [Structural differences between health insurance funds and their impact on health services research: results from the Bertelsmann Health-Care Monitor]. Gesundheitswesen Bundesverb Arzte Offentlichen Gesundheitsdienstes Ger 2012; 74: 291–7.

18. Jaunzeme J, Eberhard S, Geyer S: [How 'representative' are SHI (statutory health insurance) data? Demographic and social differences and similarities between an SHI-insured population, the population of Lower Saxony, and that of the Federal Republic of Germany using the example of the AOK in Lower Saxony]. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2013; 56: 447–54.

19. Hoffmann F, Koller D: Verschiedene Regionen, verschiedene Versichertenpopulationen? Soziodemografische und gesundheitsbezogene Unterschiede zwischen Krankenkassen. Gesundheitswesen 2017; 79: e1–9.

20. Mackenbach JP, Stirbu I, Roskam A-JR, et al.: Socioeconomic inequalities in health in 22 European countries. N Engl J Med 2008; 358: 2468–81.

21. Ohlmeier C, Langner I, Hillebrand K, et al.: Mortality in the German Pharmacoepidemiological Research Database (GePaRD) compared to national data in Germany: results from a validation study. BMC Public Health 2015; 15: 570.

22. van Raalte AA, Klüsener S, Oksuzyan A, Grigoriev P: Declining regional disparities in mortality in the context of persisting large inequalities in economic conditions: the case of Germany. Int J Epidemiol 2020; 49: 486–96.

23. Grigoriev P, Pechholdová M, Mühlichen M, Scholz RD, Klüsener S: 30 Jahre Deutsche Einheit: Errungenschaften und verbliebene Unterschiede in der Mortalitätsentwicklung nach Alter und Todesursachen. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2021; 64: 481–90.

24. Skirbekk V: Fertility trends by social status. Demogr Res 2008; 18: 145–80.

25. Dribe M, Hacker JD, Scalone F: Socioeconomic Status and Net Fertility during the Fertility Decline: A Comparative Analysis of Canada, Iceland, Sweden, Norway and the United States. Popul Stud 2014; 68: 135–49.

26. Ohlmeier C, Frick J, Prütz F, et al.: Nutzungsmöglichkeiten von Routinedaten der Gesetzlichen Krankenversicherung in der Gesundheitsberichterstattung des Bundes. Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz 2014; 57: 464–72.

27. Deutsche Institut für Medizinische Dokumentation, und Information (DIMDI): Gutachten: Daten für die Versorgungsforschung. Zugang und Nutzungsmöglichkeiten. http://www.dimdi.de/static/de/versorgungsdaten/wissenswertes/datengutachten/dimdi-sekundaerdaten-expertise.pdf (last accessed on 2017 Apr 21)

Tables

Table 1 Demographic characteristics of insurees in the InGef sample database and the total German population (as of 31.12.2018).

	InGef sample database (4.02 Mio)	Germany (83.02 Mio)
Female (%)	50.8	50.7
Age overall, years		
Mean (SD)	44.1	43.9
Median (Q1-Q3)	46 (25-62)	45 (25-62)
Age females, years		
Mean (SD)	45.5	45.3
Median (Q1-Q3)	48 (27-64)	47 (26-64)
Age males, years		
Mean (SD)	42.8	42.6
Median (Q1-Q3)	44 (24-60)	44 (24-60)
Region (%)		
Eastern Germany	16.5	19.5
Western Germany	82.3	80.5
Type of area (%)		
Rural	28.6	32.0
Urban	70.1	68.0

Figures

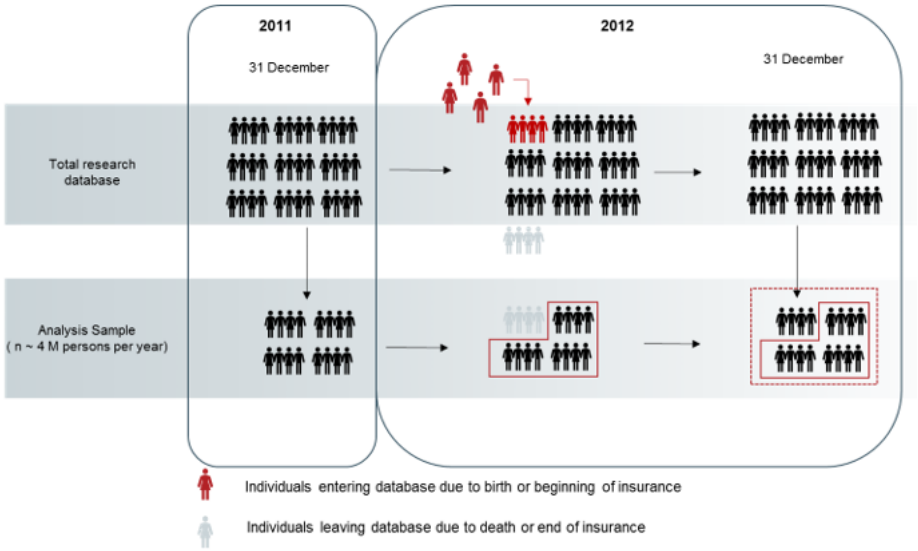

Figure 1

Schematic representation of the sampling strategy of the InGef sample database.

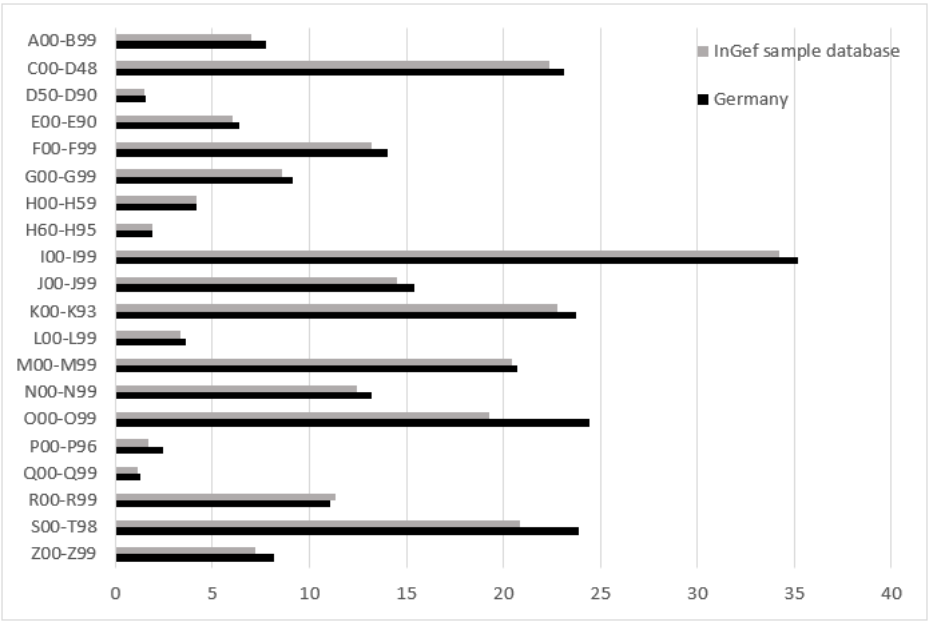

Figure 2

Hospitalization rates in the InGef database and the German population in 2019 (Rates in 2019 per 1,000 persons)

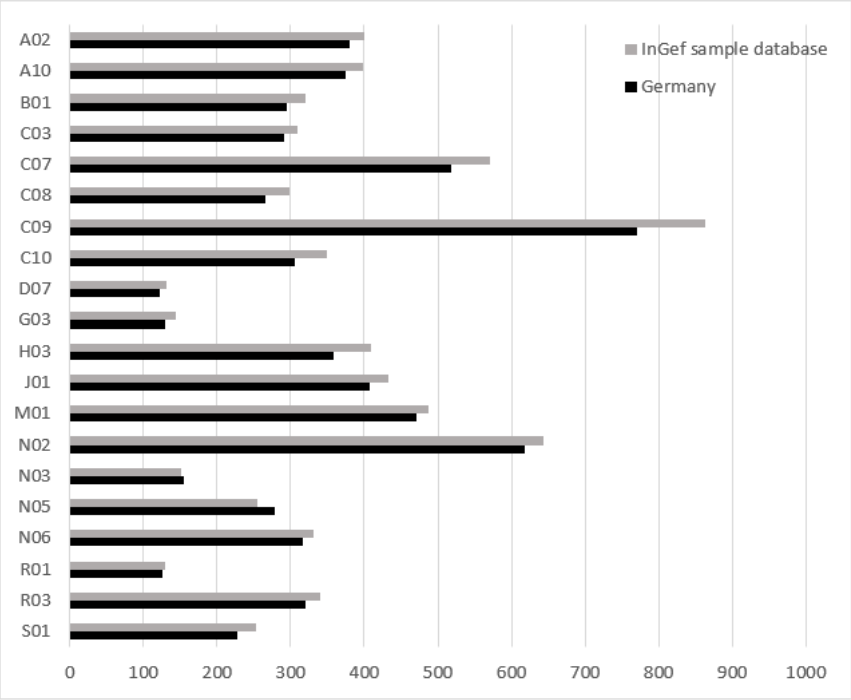

Figure 3

Drug prescription rates (ATC-codes 2nd level) in the InGef database and the German population in 2019 (Rates in 2019 per 1,000 persons)